BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 5: e73736
doi: 10.3897/biss.5.73736

OPEN ACCESS

Conference Abstract

Taxonomy Compilation & Curation Within R

Vijay Barve ‡, §

‡ Post Doctoral Researcher, Terrestrial Parasite Tracker TCN, West Lafayette, Indiana, United States of America
§ Florida Museum of Natural History, Gainesville, United States of America

Corresponding author: Vijay Barve (vijay.barve@gmail.com)

Received: 31 Aug 2021 | Published: 31 Aug 2021

Citation: Barve V (2021) Taxonomy Compilation & Curation Within R. Biodiversity Information Science and Standards 5: e73736. https://doi.org/10.3897/biss.5.73736

Abstract

Research projects in ecology or biodiversity either start with an area of study or a target species list. Working with these species lists or taxonomic lists is not as straightforward as it seems. The taxonomic names that are considered to be "standard," are surprisingly dynamic. Over time, the names keep changing with ongoing research and advancements in taxonomy. Additionally, they undergo all sorts of reorganization, such as one species being split into multiple species and/or subspecies, the grouping of multiple species into a single species, and the reclassification of species from one genus to another. Compiling a consistent target species list can be very time consuming and tricky. However it is the initial step in most research projects and needs to be completed in order to continue the research.

Advancements in biodiversity informatics are helping simplify and automate some of these tasks. There are several web services that provide taxonomic data with either a taxonomic or a geographic focus. An increasing number of experts are opening access to their carefully curated taxonomic lists. Even with the help of these services, a lot of time needs to be spent to create a working list of names that can be linked to data such as Global Biodiversity Information Facility (GBIF) mediated occurrence data.

The package *"taxotools"* (Barve 2021) provides basic taxonomic list processing functions within the R programming environment (R Core Team 2021). Even though it is a work in

© Barve V. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

progress, the functions available so far are applicable to diverse projects. The tools available can be categorized into the following broad areas:

- **Name manipulation:** A set of helper functions to check scientific names with global name resolution services like Global Names Architecture (GNA) & GBIF Name Parser, and to construct and deconstruct scientific names to and from components like genus, species and subspecific units.

- **Name matching:** Matches names either with global name services or with user-created master taxonomy lists using fuzzy matching, testing combinations of genus level synonyms, subspecies elevation to species, trying to match with higher level taxonomic entities like genus and family, and employing a user-defined lookup table to manually resolve names.

- **List processing:** Updates list fields such as unique identifiers (id), higher taxonomy and taxonomic ranks.

- **List matching:** Compares user generated lists with each other and finds differences in the two lists, then prepares the lists for merging together to form a masterlist.

- **Format conversion:** Converts taxolist to and from formats like HTML and Darwin Core (Wieczorek et al. 2021), which is useful in data exchange or checking the lists manually.

- **Name harvesting functions:** Acquires additional names from Integrated Taxonomic Information System (ITIS) and Wikipedia (taxonomy infobox).

Detailed function listings under each category are listed in Table 1.

Table 1.

List of functions in package taxotools.

Name manipulation functions	• cast_canonical: Construct canonical names • cast_cs_field: Build a character (comma) separated List within field • cast_scientificname: Cast scientific name using taxonomic fields • expand_name: Expands Scientific name • melt_canonical: Deconstruct canonical names • melt_cs_field: Generate a list melting character (comma) separated field values into multiple records • melt_scientificname: Melt scientific name into fields
Name matching	• get_accepted_names: Fetch accepted names from masterlist • check_scientific: Parse and resolve a scientific name string • get_synonyms: Fetch all synonyms for supplied names from masterlist • taxo_fuzzy_match: Use fuzzy matching to find similar names • resolve_names: Resolve canonical names against GNA

List processing functions	• compact_ids: compact id numbers • guess_taxo_rank: Guess the taxonomic rank of Scientific Name • list_higher_taxo: Get higher taxonomy data for list of names • synonymize_subspecies: Convert all subspecies into synonyms of the species • build_gen_syn: Build genus level synonyms
List matching functions	• match_lists: match two taxonomic lists • merge_lists: merge two lists of names
Format conversion functions	• DwC2taxo: Darwin Core to Taxolist format • taxo2DwC: Taxolist to Darwin Core (DwC) • taxo2doc: Taxolist to document • taxo2syn: Taxolist to Synonym list • wiki2taxo: Wikipedia list to Taxolist • syn2taxo: Synonym list to Taxolist
Name harvesting functions	• get_itis_syn: Get ITIS Synonyms for a Scientific Name • list_itis_syn: Get ITIS Synonyms for list of names • list_wiki_syn: Get Wikipedia Synonyms for list of names

This package has been effectively used in several biodiversity studies and projects like Map of Life, ButterflyNet, Terrestrial Parasite Tracker etc. It has been successfully tested on a masterlist constructed with ~1M names from World Flora Online and performs well.

The package is available on The Comprehensive R Archive Network (CRAN) [https://CRAN.R-project.org/package=taxotools] and the developmental release is on GitHub [https://github.com/vijaybarve/taxotools].

Keywords

R project, R package

Presenting author

Vijay Barve

Presented at

TDWG 2021

Conflicts of interest

None

References

- Barve V (2021) Taxotools: Tools to handle taxonomic lists. 0.0.79. R package. Release date: 2021-1-18. URL: https://doi.org/10.5281/zenodo.3934939
- R Core Team (2021) R: A language and environment for statistical computing. 4.1.0. R Foundation for Statistical computing, Vienna, Austria.. Release date: 2021-5-18. URL: https://www.R-project.org/.
- Wieczorek J, Bloom D, Guralnick R, Blum S, Dring M, et al. (2021) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1): e29715. https://doi.org/10.1371/journal.pone.0029715